A Videography Analysis Framework for Video Retrieval and Summarization

Kang Li*[1]
kangli@buffalo.edu
Sangmin Oh*[2]
sangmin.oh@kitware.com
A. G. Amitha Perera[2]
amitha.perera@kitware.com
Yun Fu[3]
raymondyunfu@gmail.com

[1] Department of CSE
State University of New York
Buffalo, NY, USA

[2] Kitware, Inc.
Clifton Park, NY, USA

[3] Department of ECE and College of CIS
Northeastern University
Boston, MA, USA

Abstract

In this work, we focus on developing features and approaches to represent and analyze videography styles in unconstrained videos. By unconstrained videos, we mean typical consumer videos with significant content complexity and diverse editing artifacts, mostly with long duration. Our approach constructs a *videography dictionary*, which is used to represent each video clip as a series of varying videography words. In addition to conventional features such as camera motion and foreground object motion, two novel features including motion correlation and scale information are introduced to characterize videography. Then, we show that unique videography signatures from different events can be automatically identified, using statistical analysis methods. For practical applications, we explore the use of videography analysis for content-based video retrieval and video summarization. We compare our approaches with other methods on a large unconstrained video dataset, and demonstrate that our approach benefits video analysis.

1 Introduction

Automatic understanding of visual content in unconstrained Internet video, such as those found on consumer video sharing sites (*e.g.*, YouTube and Metacafe), offers an interesting but very challenging task. These videos are particularly challenging because they contain very diverse content; they are captured under a variety of camera motion conditions (panning, zooming, translating); they are of highly variable length (from minutes to hours); and they are often heavily edited (*e.g.*, shot stitching and adding captions). As such, unconstrained videos are qualitatively very different and even more challenging than widely-used video datasets, such as the Hollywood dataset [5] or the YouTube Sports dataset [7], in which video clips contain fairly coherent single action occurring within a short duration. For example, some wedding videos from video sharing websites are more than an hour long and they are produced by stitching shots recorded separately across the entire wedding event. Each shot contains fairly different content, such as a panning camera capturing a party room filled with dancing guests, a series of stitched shots of each guest individually congratulating the

* Indicates equal contributions.
http://dx.doi.org/10.5244/C.26.126

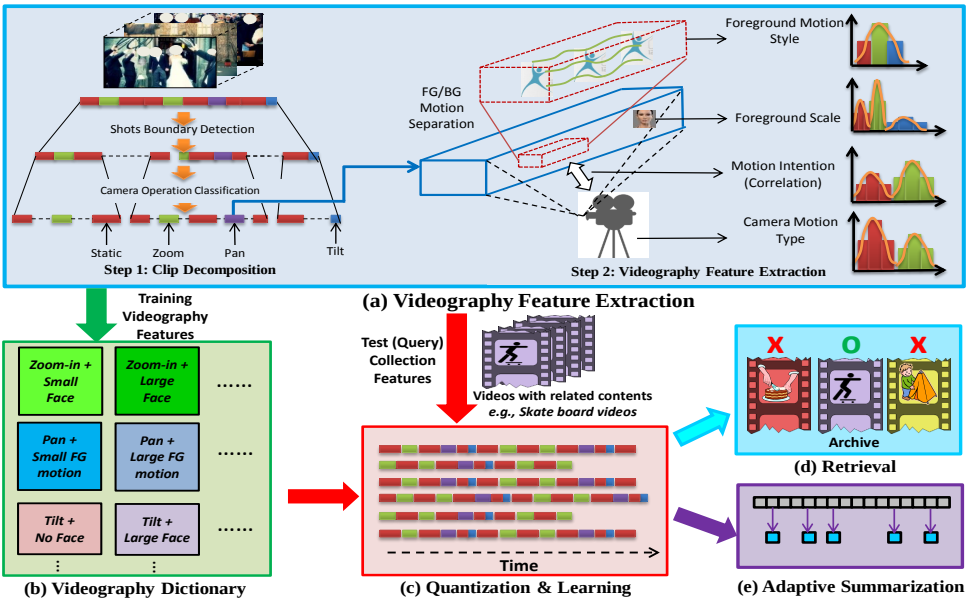


Figure 1: Framework for videography analysis and applications for unconstrained videos. See text for details.

wedding, or a shot that zooms in on the bride and groom. On the other hand, other wedding videos may be only minutes long, and only contain shots of the key events of the ceremony.

In this work, we present an approach for *unsupervised videography analysis* for this type of unconstrained video. Intuitively, each videography can be understood as a camera director's direction on a movie script, *e.g.*, "capture the running actress by panning the camera, to have her face appear at 20 percent size of the video". The idea is that different classes of video content will have different videography styles—the videography style of a wedding video should be different from a sports video—and so, the videography style should provide a valuable signal for automated content analysis. In this paper, we demonstrate the value of videography analysis for video retrieval by event class and for video summarization.

In our approach, we assume that there are diverse videography styles in unconstrained videos, which are discovered as a *videography dictionary* via unsupervised clustering on proposed features. Then, a video clip can be represented as a series of segments with varying videography words. For the underlying videography features, we extend conventional features such as camera motion and foreground (FG) object motion [3, 11, 15, 19, 25] by incorporating two novel features: *motion correlation* and *scale* information (see Sec. 3). To the best of our knowledge, our work is the first to address the explicit learning of a videography dictionary based on such a rich set of features beyond simple camera motions.

The overview of our proposed approaches is illustrated in Fig. 1. We first (a) extract the videography features by decomposing the video into segments based on camera motion-

We thank Jingchen Liu, Ben Miller, and Scott McCloskey for sharing GIST and face detection results. This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.

derived "shot boundaries", separating foreground/background motion within each segment, and computing a series of features (as illustrated in Step 2 and described in Sec. 3). Then we (b) cluster these features to develop a videography dictionary, and (c) quantize the segments into videography style words and learn the relationship between the style words and events. This is used for (d) video retrieval, and (e) to help content-adaptive video summarization.

For retrieval, we compare our approach with alternative methods on a large TRECVID multimedia event detection (MED) '11 video dataset [1] across 15 different diverse query collections, and show that the videography style does indeed add complementary information (Sec. 5). In addition, our adaptive summarization approach is different from the existing body of work relying on fixed rules (*e.g.*, [25]) in that our system optimizes summarization process to highlight the unique content of the given test videos (Sec. 6).

2 Related Work

The idea of representing videos as a series of segments based on motion and/or appearance characteristics has been explored to some extent, either as part of integrated systems [19, 22, 25] or on its own [11, 19]. Most systems, including this work, incorporate two main low-level processing steps: (a) shot boundary detection [6, 18, 23], which is to find the boundaries between stitched shots, and (b) camera motion estimation within shots [2, 11, 14, 19, 24, 25] to further decompose shots into finer sub-shot units based on evolving camera motion types.

It is worth noting that we incorporate existing state-of-the-art methods as part of our feature extraction module, and focus on (a) developing novel techniques to enable high-level videography analysis and (b) its application for retrieval and summarization based on noisy videography quantization as intermediate representations. Shot boundary detection is believed to be largely solved [18]; we adopt [23]. For background (BG) camera motion estimation, we extend [14, 24] to estimate three P/T/Z camera motion parameters from KLT tracks while simultaneously separating the tracks into FG/BG groups. We found that other approaches for FG/BG separation such as [3, 13] are unsatisfactory for unconstrained videos, possibly due to the complex geometric scene structure in our data.

In terms of videography modeling, the methods closest to our work are [22, 25]. In [25], a system capable of both summarization and retrieval was presented. The system is mostly based on hand-tuned distance metrics and rules to classify shots and videos into semantic categories, based on multiple features with heavy emphasis on appearance (*e.g.*, color and texture), and a few others such as simple camera motion primitives (S/P/T/Z). In our retrieval experiments (Sec. 5), we compare our new features with these simpler 4 types of camera motion primitives. It is worth noting that our work presents results primarily based on motion information without relying on appearance matching, hence, provides a clearer understanding on the promise of motion-based videography modeling alone for high-level tasks. Additionally, since our approach is learning-based, the heavy burden to tune system parameters is alleviated. In [22], the authors present seven self-defined videography styles common in commercial movies, which are classified per shot based on features such as motion, appearance, and FG/BG separation; the videography quantization is based on supervised learning, and its use for summarization or retrieval is not studied. In contrast, our approach is unsupervised and does not require manually labeled training data for sub-shot classification, and hence can scale up for unconstrained videos with more complex videography styles beyond commercial movies.

For video retrieval based on videography, other than the above-mentioned related work in [25], [15] used simple average profiles of FG/BG motion magnitude as features. In [19], the correlation between different categories of sports videos and camera motion types (*e.g.*,

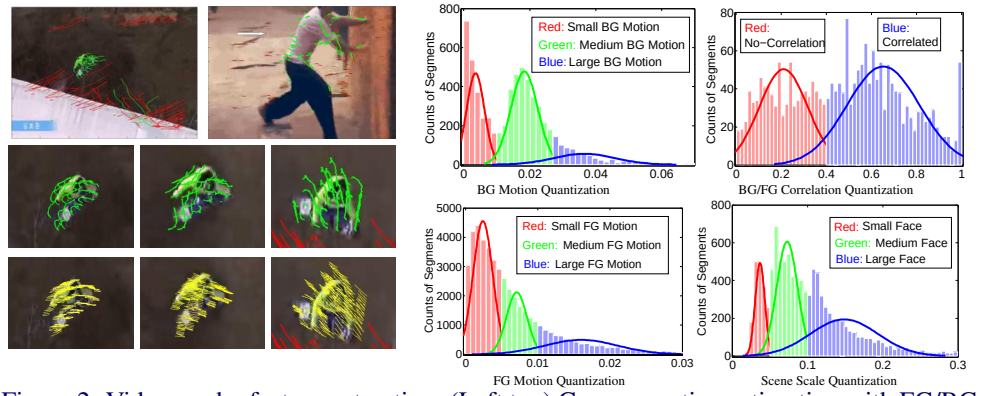


Figure 2: Videography feature extraction. (Left top) Camera motion estimation with FG/BG separation. (Left middle and bottom) Original FG motion (green) is corrected (yellow). (Right) Distribution of extracted videography features, and a clustering-based quantization.

S/P/T/Z) and their transitions were studied, but without a notion for retrieval.

Video summarization that has been well studied in multimedia community [9] is formulated as key frame extraction problem where change detection is commonly used based on appearance features such as color [22]. Different approaches which incorporate overall camera motion include [2, 25]. However, both works adopted fixed rules for all videos.

3 Videography Features

For every input video, our approach applies two main processing steps to extract videography features, as illustrated in Fig. 1(a). First, a two-level motion analysis is conducted to decompose long clips into sequences of segments with coherent motion types (S/P/T/Z). Second, multiple features related to motion and scale patterns are measured from every segment, which are used to characterize videography. For both steps, we utilize densely computed KLT tracks [17] over the entire clips as main basis for the derived features.

For the two-level decomposition, we adopt existing state-of-the-art methods, as mentioned in Sec. 2. In the first phase, we use a shot boundary detection (SBD) algorithm which relies on the birth and death ratio of KLT tracks [23]. In detail, we developed two SBD modules, each one for two different styles of boundaries, namely: *Cut* (simple abrupt transition) and *Fade-Out-In* (common gradual transition), which account for majority of boundaries in videos. On labeled test data of 153 shot boundaries, the precision and recall are 0.95 and 0.98 for *Cut*, and 0.63 and 0.75 for *Fade-Out-In*, which are fairly good results.

Then, the second phase decomposes each shot further into sub-segments based on four camera motion types (S/P/T/Z). For unconstrained videos, camera motion estimation is challenging due to the complex interplay between the (apparent) motion of background (BG) and foreground (FG) objects, which need to be separated to yield accurate results. We adopt [12, 24] because of its proven performance on unconstrained videos and its advantage of solving FG/BG separation simultaneously. As a result, KLT tracks are grouped into BG or FG, where BG group accounts for tracks mostly induced by camera motion and FG group as outliers from BG. Furthermore, to capture motion characteristics of FG objects accurately, FG tracks are motion-corrected by subtracting average BG motion. These are illustrated in Fig. 2(Left). Although FG/BG separation results are not perfect, the portion of mis-classified tracks is usually small, hence, unlikely to undermine the overall videography analysis.

Once segments are obtained, a set of videography features are extracted from every segment. In this work, we focus on visual features related to *motion* and *scale*: (1) camera motion type (S/P/T/Z), (2) FG and (3) BG motion, (4) correlations between FG/BG motion, and (5) the scale of foreground. For FG and BG motion, the average motion within a segment is normalized w.r.t. the video width, to cope with video clips with varying sizes. Our novel FG/BG correlation feature is motivated by the fact that similar camera motion may be invoked by different intentions, *e.g.*, tracking or simply switch of focus. The magnitudes of FG/BG correlation are measured by the normalized sum of inner product between FG tracks and average BG motion. We also include scales of FG objects as another distinctive feature for videography. For example, clips with close-up shots of faces are very different from clips which contain far-away shots of pedestrians. Because the estimation of scale is a very challenging problem, in this work, we used the bounding box sizes of face detections produced by off-the-shelf systems (*e.g.*, [21]) as a proxy for scale estimates. In detail, average face size within a segment (normalized by the video height) is used to represent the scale. For example, face scale of 0.2 indicates that the average size of faces occupies about 20 percent of the image height. It is worth noting that, there are alternative approaches for scale estimation by solving depth [16] or 3D geometry [20]. However, applying such methods for unconstrained videos is beyond the scope of this work, and is left for future work.

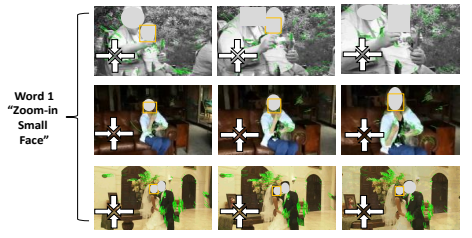
For our experiments, we extracted the above-mentioned videography features from a training video dataset, which consists of roughly 2000 unconstrained videos (~80 hours total), where 29 segments are found per clip on average. The overall distribution of the extracted features are shown in Fig. 2(Right), where the multi-modal characteristics in most videography features (except FG motion) can be observed. Such patterns indicate that there are indeed regularized videography patterns in videos.

4 Videography Dictionary and Analysis

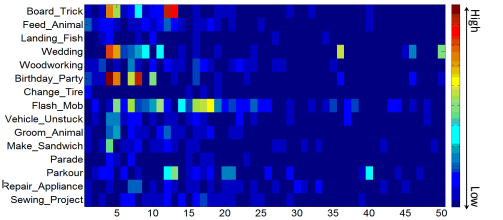
Once videography features are obtained from segments, they are grouped to form videography dictionary (VD) shown in Fig. 1(b). The computed VD will be used to quantize video clips into sequences of videography words (VWs), as shown in Fig. 1(c).

We have explored two different methods for developing the dictionary: (1) concatenated and (2) joint learning. In the first *concatenated* learning, each feature dimension is quantized individually, then, are concatenated to form VD in a combinatoric manner. Straightforwardly, the first feature dimension of camera motion type has four quantization values of S/P/T/Z. We quantize the remaining features individually, based on an empirical analysis of the data on the training set. As illustrated in Fig. 2(right), the BG and FG motion is each quantized into *small/medium/large*; the FG/BG correlation into *correlation* or *no-correlation*; and the scale into *no-face/small/medium/large*. The video words are then formed by concatenating these values. This creates $4 \times 3 \times 3 \times 2 \times 4 = 288$ possible video words.

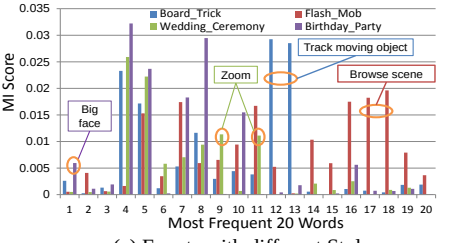
Our analysis of the distribution of the resulting VD shows that, interestingly, only ~40% of the words are actually observed in the data, indicating that only a subset of combinations of feature quantizations are present, *e.g.*, a combination such as zoom-in, large FG and BG motion, no correlation, and large scale actually does not appear. Furthermore, if we eliminate rare words which have fewer than ten occurrences, we are left with only 82 unique videography words, over a dataset of 80 hours of unconstrained video. Such observation provides an insight that there are fairly regularized patterns in how people capture videos, regardless of content. To the best of our knowledge, this is the first study that provides automated analysis on characteristics of videography styles on unconstrained Internet videos.



(a) Videography word examples



(b) Mutual info. between events and words



(c) Events with different Styles

Figure 3: (a) Videography word examples. (b) Mutual information between different event classes and most frequent 50 VWs. (c) Qualitative analysis on 4 event classes.

In the second method of *joint learning* for developing the dictionary, we again quantize the motion type into the same four values (S/P/T/Z). However, for each motion type, we perform K-means clustering on the remaining four-dimension continuous vector space formed by concatenating the four raw feature types (FG motion, BG motion, amount of correlation, size of face). In our experiments, we chose K=30, which yields $4 \times 30 = 120$ video words. We used a smaller number of clusters because of the observation that many of the video words from the first method were actually not used.

Once VDs are obtained, we can examine their accuracy as a macro feature type by examining the sample video segments in each word cluster. Example segments belonging to two sample videography word clusters are shown in Fig. 3(a), along with the detected visual features overlayed on images to show more details, including camera motion (left bottom arrows), compensated FG motion (green tracks), and face detections (orange boxes)[1]. The textual descriptions of both words were produced manually, by looking at both the feature vector values and the grouped segments. It can be observed that segments with highly related content are successfully grouped into the same VWs. In particular, it is worth noting that in the second example, similar segments are grouped together correctly, even though faces are not detected due to the challenging imaging conditions. We have manually examined 10 VWs by drawing 30 segment samples each and concluded that, on average, 88% of segments from the same VWs show perceptually identical videography.

We also conducted analysis on the correlations between VWs and particular visual content, so called *events*. By events, we mean semantic content classes captured in videos, such as *Flash mob* or *Birthday party* (defined further in [11]). This notion of analyzing or learning about videography of videos containing the same events is illustrated in Fig. 3(b,c). Specifically, we measured the mutual information (MI) between each word and each event. A high MI score indicates that a word is discriminative for the corresponding event. Our results are summarized in Fig. 3(b) where MI between every event and top 50 most frequent VWs are shown. It can be observed that, for a particular event, there are certain *signature* VWs. More

[1] In this work, faces are intentionally occluded in this figure for privacy.

detailed analysis is shown for four event types and top 20 words, in Fig. 3(c). In particular, this analysis provides insight on how different events are captured with different styles. For example, it shows that event *Board trick* has a strong style of *tracking moving object*; event *Flash mob* has a strong style of *browsing scenes*; event *Wedding ceremony* shows frequent *zooming*; and event *Birthday party* shows frequent *facial close-up*. This observation on discriminative correlations suggests that videography analysis can actually be used for challenging tasks such as retrieval (Sec. 5) and summarization (Sec. 6).

5 Application for Video Retrieval

In this section, we present our approach and experimental results for videography-based video retrieval. In detail, we computed videography word bag-of-word (VW-BoW) representations, where per-clip unigram features are built from sequence of VWs (regardless of temporal ordering), for every clip. The goals are to examine (1) how well the proposed VW-BoW feature can perform in retrieval tasks by itself, compared to other alternatives and with detailed studies on contribution of each videography feature component, and (2) whether our approach offers a useful modality to capture characteristics of video belonging to high-level event classes, in comparison to other macro-level features such as GIST [17].

For dataset, we use TRECVID 2011 multimedia event detection (MED) corpus [1] as our data, due to its large size, realistic content variability, and existing clip-level annotations for 15 different event classes. Both the scale and complexity of the dataset are beyond the widely-used datasets [6, 7]. Clips are frequently captured in unconstrained lighting and camera motion conditions, exhibiting diverse degrees of encoding artifacts and severe background clutter, and heavily edited by owners using shot stitching, caption embedding, etc. For training data, we use "Part-1 training data" (called event kits), which consists of videos from 15 different event classes of 137 clips per class on average (total 2061 clips) with average duration of 4.2 minutes. From these training data, our VDs are computed by selecting the best run out of 100 K-means clustering, and later used for test data. The 15 event types are enlisted in the caption of Fig. 4, with events frequently exhibiting complex camera motion marked in bold faces. For test data, MED corpus provides two different subsets, "Part-1 DEV-T" for the first 5 event classes, and "MED11TEST" for the remaining 10 event classes, with 4292 and 32061 total clips respectively. Both test datasets contain large amount of negative clips which do not belong to any of the target event classes, consequently, they serve as realistic test-bed for retrieval experiments. The positive examples in the two test datasets only constitute 2.34% and 0.37% on average per class respectively.

Our retrieval experiments are conducted using one-vs-all SVM classifiers, parameters of which are tuned via cross-validation. The overall results are summarized in Fig. 4 and Table 1, where several experiments are conducted[2]. As performance metrics, average precision (AP) is used. It is worth noting that APs for E06-E15 are lower than E01-E05, because the relative ratio of negative samples in the test dataset for E06-E15 is about 10 times higher. In detail, *Chance* denotes random retrieval and *PTZ* denotes the use of four-dimensional BoWs of discrete camera motion types only (*e.g.*, S/P/T/Z) without detailed videography features, as comparative methods [19, 25]. The variations of our approaches are marked using abbreviations where J and C denote joint or concatenated VD learning, described in Sec. 4. Additionally, B, F, C, S indicate the inclusion of BG motion, FG motion, BG/FG correlation, and scale respectively, during VD learning. These experiments have been conducted to examine the usefulness of each videography feature for retrieval. The minus sign '$-$' indicates

[2]Detailed numerical values of all experimental results can be found in supplemantal materials.

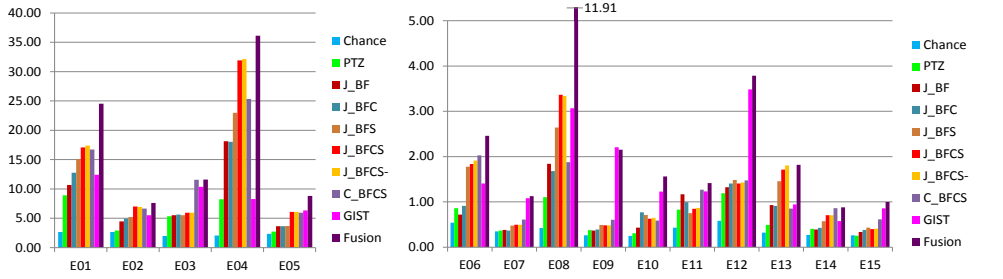


Figure 4: Average Precision (%) of video retrieval results on MED corpus, for 15 events: ***(E01) Board trick***, *(E02) Feeding animal, (E03) Fishing,* ***(E04) Wedding***, *(E05) Working wood project,* ***(E06) Birthday party***, *(E07) Change vehicle tire,* ***(E08) Flash mob***, *(E09) Getting vehicle unstuck, (E10) Groom animal, (E11) Make sandwich,* ***(E12) Parade***, ***(E13) Parkour***, *(E14), Repair appliance, and (E15) Sewing project.*

Table 1: Mean average precision (%) of video retrieval results on MED corpus, for two separate test datasets of events (left) 1-5 and (right) 6-15 respectively. Fusion results are obtained by combining J_BFCS and GIST. The results with dynamic events only are marked with (D), which include events: E01, E04, E06, E08, E12, and E13.

mAP	Chance	PTZ	J_BFCS	J_BFCS(D)	GIST	GIST(D)	Fusion	Fusion(D)
E01-E05	2.34	5.63	13.61	24.50	8.57	10.34	17.74	30.35
E06-E15	0.37	0.62	1.19	2.08	1.61	2.22	2.81	4.99

that the VD has been pruned by filtering out VWs with low MI scores per event type. For all the experiments with BoW-type features, histogram intersection kernel (HIK) was used for SVM training and testing. In addition, *GIST* shows the results using GIST features [12] with linear SVMs. Because GIST is a per-image feature, GIST features are computed on frames extracted from labeled video clips. Then, one-vs-all SVMs were trained on image features using clip labels. For testing, SVMs are applied on extracted images, then, scores were averaged to produce a clip-level score. Apparently, VWs and GIST capture very distinct signals from data. Accordingly, in the experiment marked as *Fusion*, we have further explored whether fusion of two modalities can lead to further improvement, which will show whether these two feature types are complementary. For fusion, we have used the approach of "late fusion" (*e.g.*, [4]) where we have used the weighted sum of two classifiers as the fusion score. Among VW-based approaches, *J_BFCS* was used because it has been shown to provide best performance, and weights were determined by cross validation where equal weights of <0.5, 0.5> were found to be best.

Overall, we can observe that VWs clearly provide advantage over the conventional simpler alternative of using camera types only, *i.e.*, *PTZ*. From Fig. 4, it can also be observed that every videography feature contributes towards improving performance. Between joint and concatenated VD learning, joint learning shows superior performance overall, possibly due to the data-driven construction of the dictionary which avoids many empty (or rare) VWs in concatenated learning. However, pruning VWs by MI scores does not seem to necessarily boost performance. Table 1 shows mean average precision (mAP) for key experiments in Fig. 4 on two test datasets. It can be observed that motion-based macro feature such as videography can outperform GIST for E01-E05 in "Part-1 DEV-T" set, and E06, E08, E11, E13, E14 in "MED11TEST" set. More importantly, the fusion results are much better than either approach, indicating that two feature types are complimentary. Table 1 also shows

mAPs for dynamic events only, where we observe big boost in performance for VWs. Interestingly, the event classes which show clear discriminative correlation with VWs in Fig. 3(b) are dynamic events, and they also show more advantage when VWs are used for retrieval.

6 Application for Video Summarization

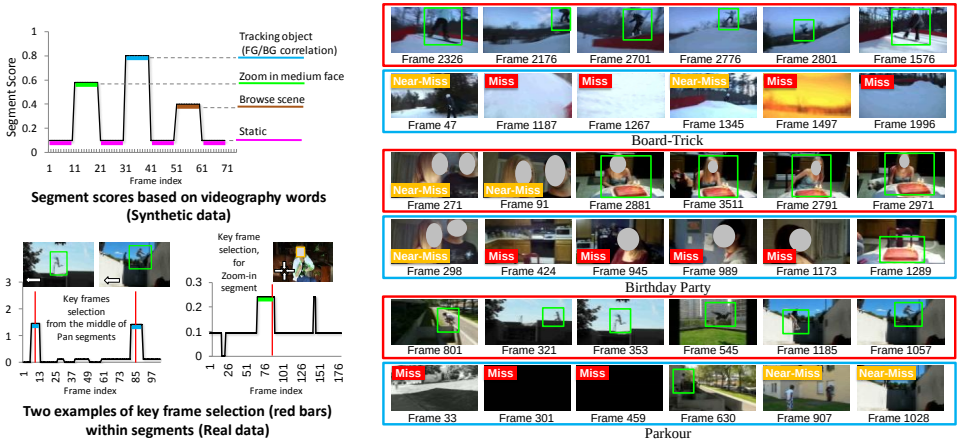


Figure 5: Videography-aware adaptive summarization. (Left) Segment scores are based on MIs of corresponding VWs. Frames are selected at designated relative location within segments. (Right) Three summarization results by this work (red rows) and baseline (blue rows). Detected FG regions (green) and human judgements on relevance of key frames (good:none, near-miss: yellow, miss: red) to associated events are marked on each image.

In this section, we present our videography-aware adaptive summarization method, which is designed to highlight the segments with distinctive videography styles for particular events. Our novel insight is that identification of segments from videos where cameramen are systematically exhibiting distinctive videography styles for particular events will provide unique summarization, assuming that such segments are strongly correlated with the major region of interest. While many works deliberately avoid the use of segments with motion due to complexity, *e.g.*, [10], such segments can be indeed crucial to characterize dynamic contents in videos exhibiting frequent camera motion, frequently recorded by mobile devices.

In our approach, frames are extracted by two step procedures, as illustrated in Fig. 5(Left). First, key segments are selected based on segment scores, with optional weighted sampling scheme in case there are more number of segments than the desired number of key frames. For segment scores, MI scores have been used[3]. Then, key frames are extracted, one per selected segment. In particular, our novel innovation is that frames are designed to be extracted from different relative location within each segment based on their videography. Two different types of key frame selection mechanisms were used: frames are selected (1) in the middle of segments when videography is either stationary or indicates FG/BG correlation (to capture peak of FG motion), and (2) at either end of segments when videography indicates P/T/Z without FG/BG correlation (to capture the destination of shifting attention).

Qualitative summarization results are shown in Fig. 5(Right), where frames extracted from same videos by our proposed method (red rows) and a conventional baseline (blue rows) are compared, for three different event classes. The results of the baseline method were

[3] Without event labels, term frequency inverse document frequency (tf-idf) scores [8] can be used instead.

obtained by extracting frames with highest scores based on color histogram changes, which is very common [9]. It can be observed that our method is very effective in identifying unique contents from clips. In particular, most extracted frames contain important visual moments when the FG people are at the peak of their action or camera focus, such as skilled jumps or before blowing a birthday cake candle. On the other hand, results by the baseline tend to include frames that just exhibit strong changing background or even black frames around the captions inserted by users. Overall, we observe that the proposed method can generate good visual summaries, especially for clips which contain complex camera motions.

7 Conclusion

We have presented our framework for videography learning and analysis, and its application for video summarization and retrieval. The introduced features and data-driven VD learning helps to identify characteristic videography among videos from same events. Our experiments show that meaningful summarization and retrieval results can be obtained using videography. Fusion results indicate that videography features capture unique aspects of videos and can be jointly used with other features to improve retrieval substantially.

References

[1] 2011 TRECVID MED Evaluation Plan v3.0. http://www.nist.gov/itl/iad/mig/upload/MED11-EvalPlan-V03-20110801a.pdf.

[2] G. Abdollahian, C. Taskiran, Zygmunt Pizlo, and E. J. Delp. Camera motion-based analysis of user generated video. *IEEE Transaction on Multimedia*, 12(1):28–41, 2010.

[3] Nazli Ikizler-Cinbis and Stan Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *ECCV*, 2010.

[4] Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *Proceedings of ACM International Conference on Multimedia Retrieval (ICMR)*, 2011.

[5] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[6] Rainer Lienhart. Comparison of automatic shot boundary detection algorithms. In *SPIE Conference in Storage and Retrieval for Image and Video Databases*, pages 290–301, 1999.

[7] Jingen Liu, Jiebo Luo, and Mubarak Shah. Recognizing realistic actions from videos ïn the wild.̈ In *CVPR*, 2009.

[8] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[9] Arthur G. Money and Harry Agius. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Comunication and Image Represention*, 19(2):121–143, February 2008. ISSN 1047-3203.

[10] Do Hang Nga and Keiji Yanai. Automatic construction of an action video shot database using web videos. In *IEEE ICCV*, 2011.

[11] Chong-Wah Ngo, Ting-Chuen Pong, Hong-Jiang Zhang, and Roland T. Chin. Motion Characterization by Temporal Slices Analysis. In *CVPR*, 2000.

[12] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal on Computer Vision (IJCV)*, 42 (3):145–175, 2001.

[13] Shankar Rao, Roberto Tron, René Vidal, and Yi Ma. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *IEEE PAMI*, 32(10):1832–1845, 2010.

[14] Gagan B. Rath and Anamitra Makur. Iterative least squares and compression based estimations for a four-parameter linear global motion model and global motion compensation. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 9(7):1075–1099, October 1999.

[15] M. J. Roach, J. S. D. Mason, and M.Pawlewski. Video genre classification using dynamics. In *IEEE ICASSP*, 2001.

[16] Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng. 3-d depth reconstruction from a single still image. *International Journal of Computer Vision (IJCV)*, 76:2007, 2007.

[17] Jianbo Shi and Carlo Tomasi. Good features to track. In *CVPR*, 1994.

[18] Alan F. Smeaton, Paul Over, and Aiden R. Doherty. Video Shot Boundary Detection: Seven Years of TRECVid Activity. *CVIU*, 114(4):411–418, 2010.

[19] S. Takagi, S. Hattori, K. Yokoyama, A. Kodate, and H. Tominaga. Sports video categorizing method using camera motion parameters. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2003.

[20] Lorenzo Torresani, Aaron Hertzmann, and Christoph Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *PAMI*, 30(5):878–892, 2008.

[21] Paul A. Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[22] Hee Lin Wang and Loong Fah Cheong. Taxonomy of directing semantics for film shot classification. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 19(10):1529–1542, 2009.

[23] Anthony Whitehead, Prosenjit Bose, and Robert Laganiere. Feature Based Cut Detection with Automatic Threshold Selection. In *International Conference on Image and Video Retreival (CIVR)*, 2004.

[24] J. Yuan, H. Wang, and et.al. Tsinghua University at TRECVID 2005. In *TRECVID workshop*, 2005.

[25] Xingquan Zhu, Ahmed K. Elmagarmid, Xiangyang Xue, Lide Wu, and Ann Christine Catlin. InsightVideo: Towards hierarchical video content organization for efficient browsing, summarization and retrieval. *IEEE Transactions on Multimedia*, 7(4):648–666, 2005.